

Linear_regression_report

Team: Who's there

Introduction

The aim of the project is to be able to find a model that best predict the relationship between:

- the number of people counted with the survey for a given class at a particular hour
- Wi-fi log counted in that room at that hour

This will validate whether Wi-Fi log is a good predictor for estimating room occupancy.

We first set out if the relationship between these two variables was linear. To do so we ran a linear regression.

Below we describe in detail step by step how the analysis was performed.

First of all, we set up a connection to the database, using the following code:

```
connection <- dbConnect(MySQL(),user="root", password="",dbname="who_there_db", host="localhost")
```

Then we made a query to the database, in order to get all the ground truth data collected in room B.002, B.004 and B.006 from 9:00 a.m. to 17:00 p.m. and the correspondent Wi-Fi Log measured in that time frame and rooms.

The dataset created had in total 216 rows, which are sufficient for running this analysis.

As a **target features** for our linear regression we decided to use the number of people present in the room, calculated by multiplying the percentage of the room with the capacity of the room.

As dependent features we considered Wi-Fi logs, which were summarised either as average of the logs counted for each room and for each hour or as maximum of the logs measured for each room and for each hour.

Together with the Wi-Fi log, we included in the data set the following features:

- **Date**, which we did not use in this analysis, because they just cover 2 weeks of November, but for future analyses they can be used to group observations by seasons or semesters or to find seasonal trends for time series analyses.
- **Room**, which is a categorical variable with 3 levels: B002, B003, B004
- **Time**, which will be explored either as continuous variable and as categorical. For exploring time also as a categorical variable, we binned the time in 4 ranges: early morning (9:00-11:00), late morning (11:00-13:00), early afternoon (13:00-15:00) and late afternoon (15:00-17:00). This allowed us to see if the Wi-Fi log accuracy was changing during the day. For example, it is more likely that all the electronic devices are fully powered early in the morning and consequently the Wi-fi log data can be more accurate or overestimating the occupancy of the room (i.e. more than one device per person). On the contrary in the afternoon, the devices may be more likely to be out of battery and it is possible that there are less devices in the room.
- **Module**, which we are not going to include it in the analysis because the majority of the module present are for computer science. For future analyses it will be possible to explore if the accuracy of Wi-fi log in predicting the occupancy change across the courses. Science courses (especially computer science courses) will be more likely to use electronic devices during lectures than art students.

- **Course level**, which can indicate us whether electronic devices will be less used during different course levels. For example, first and second level courses can be more theoretical than the upper levels. Therefore, we might expect fewer connections. However, undergraduates might be more distracted during lectures and look at their phones. This will result in an increase of connections in that hour.
- **Tutorial**, which can affect the number of logged people. First of all, because tutorial divided the room in 2 and therefore there will be measured less people than expected.
- **Double_module**, categorical variable indicating whether in the class there are more than one module, increasing the number of people expected in the room.
- **class_went_ahead**, categorical variable indicating whether in the class went ahead to check for false positive.

The resulting data set is printed below:

```
head(AnalysisTable)
```

```
##   Room      Date Time      Module Course_Level Tutorial Double_module
## 1    1 2015-11-03   9          0          0         0          0
## 2    1 2015-11-04   9 COMP30190          3         0          0
## 3    1 2015-11-05   9          0          0         0          0
## 4    1 2015-11-06   9 COMP30220          3         0          0
## 5    1 2015-11-09   9 COMP30190          3         0          0
## 6    1 2015-11-10   9          0          0         0          0
##   Class_went_ahead Capacity Percentage_room_full Wifi_Average_logs
## 1                  1          90              0.00          4.7500
## 2                  1          90              0.25          13.4545
## 3                  1          90              0.00           6.8333
## 4                  1          90              0.00           2.4167
## 5                  1          90              0.25          14.7273
## 6                  1          90              0.00           2.2727
##   Wifi_Max_logs Survey_occupancy   Factor_Time
## 1              21              0.0 Early Morning
## 2              15             22.5 Early Morning
## 3              29              0.0 Early Morning
## 4               3              0.0 Early Morning
## 5              18             22.5 Early Morning
## 6              14              0.0 Early Morning
```

DATA QUALITY REPORT

Before running any analyses, we carried out the data quality report to check for any issue related with the variables (e.g. outliers, skewed distribution, NaN values) and we displayed the solutions that we are willing to implemented to solve them in the data quality plan table (see below).

Initially we set all the categorical variables as factors and then we print the descriptive statistics for all the features.

```
##   Room      Date      Time      Module      Course_Level
## 1:72 Length:216      Min.    : 9.00 Length:216      0:59
## 2:72 Class :character 1st Qu.:10.75 Class :character 1:14
## 3:72 Mode  :character Median :12.50 Mode  :character 2:23
##                               Mean   :12.50                               3:76
```

```

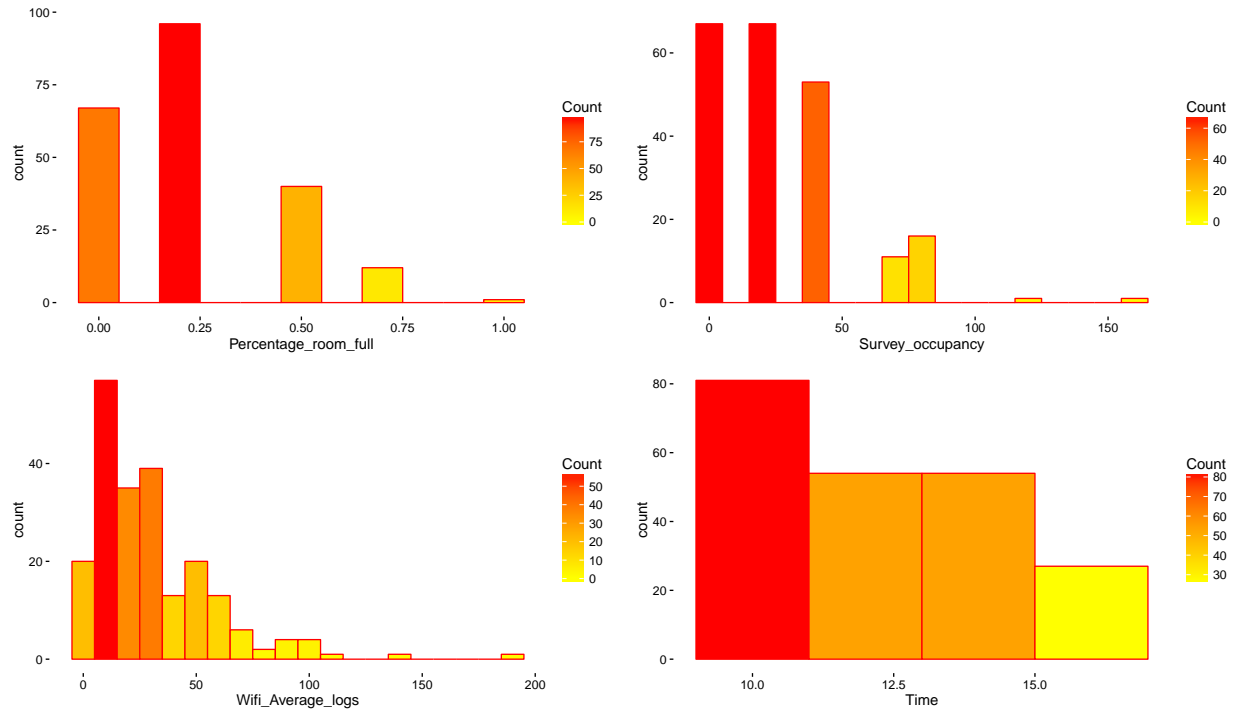
##          3rd Qu.:14.25          4:40
##          Max.    :16.00          5: 4
## Tutorial Double_module Class_went_ahead Capacity
## 0:210    0:210        0: 22      Min.   : 90.0
## 1: 6      1: 6        1:194      1st Qu.: 90.0
##                                     Median : 90.0
##                                     Mean    :113.3
##                                     3rd Qu.:160.0
##                                     Max.    :160.0
## Percentage_room_full Wifi_Average_logs Wifi_Max_logs Survey_occupancy
## Min.    :0.00        Min.    : 0.00  Min.    : 0.00  Min.    : 0.00
## 1st Qu.:0.00        1st Qu.: 11.61  1st Qu.: 18.75  1st Qu.: 0.00
## Median :0.25        Median : 23.67  Median : 32.50  Median : 22.50
## Mean    :0.25        Mean    : 30.33  Mean    : 40.01  Mean    : 28.01
## 3rd Qu.:0.25        3rd Qu.: 43.88  3rd Qu.: 55.25  3rd Qu.: 40.00
## Max.    :1.00        Max.    :192.92  Max.    :230.00  Max.    :160.00
##          Factor_Time
## Early Morning  :54
## Late Morning   :54
## Early Afternoon:54
## Late Afternoon :54
##
##

```

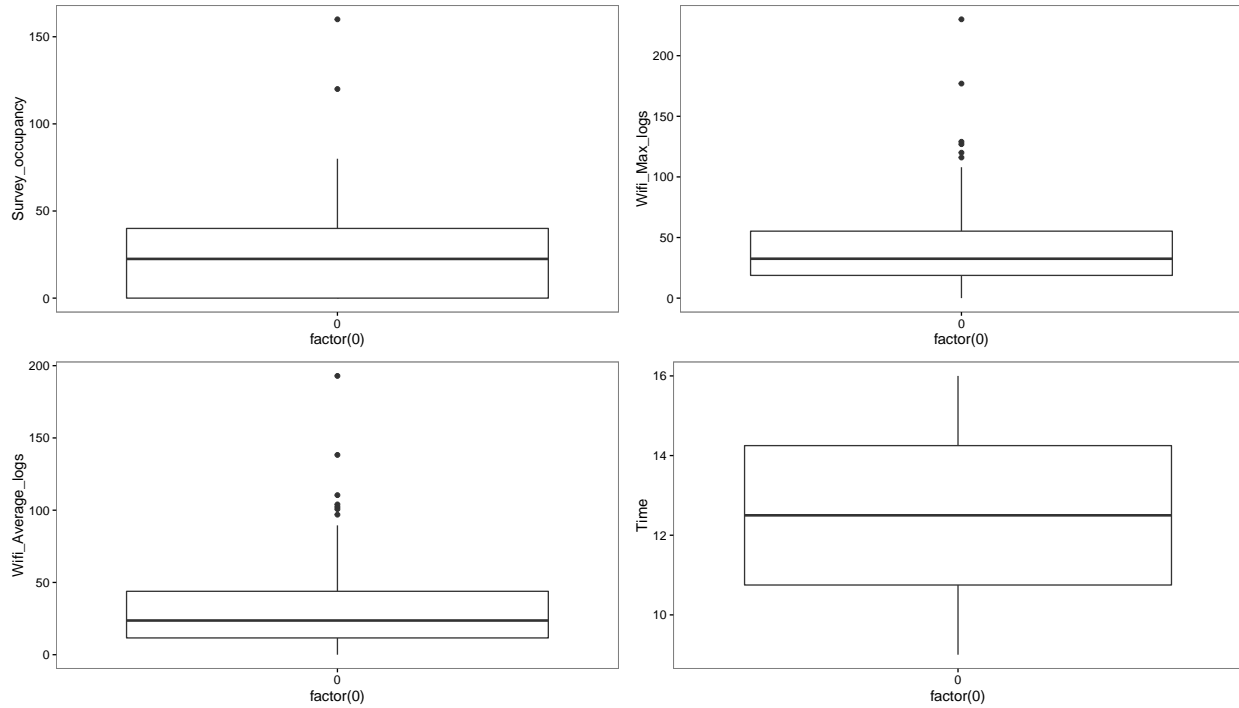
From the descriptive statistic we can see that the data set does not contain any NaN values, but we could notice several issues. For example the features Tutorials, Double_model and class_went_ahead were not even distributed across the 2 levels of the variables. Therefore, we decided to discard those features from the analyses, because they will not be informative. Furthermore, for the variables Wifi_Average_clients, Wifi_Max_clients and Survey_occupancy it seems that there were few outliers, since their median is lower than the mean and the max values were far higher than the mean values. We explored this issue with histograms and boxplots.

Exploratory graphs

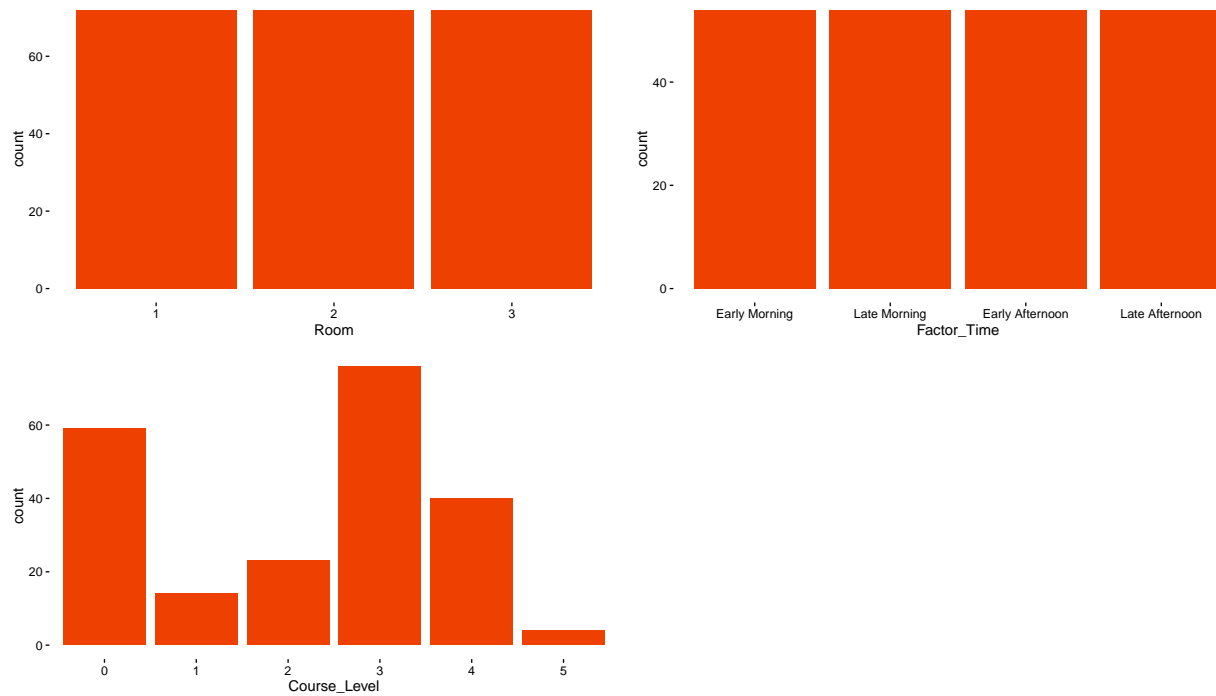
Histograms



From the histograms we could see that the distribution of the feature Wifi Maximum_client (i.e. the Maximum number of devices logged in one hour lecture) was skewed to the left, indicating that no more than 40 people attended the majority of the lectures. Furthermore, we could see that there were potential outliers. Similar patterns were observed for the feature Wifi Average_clients. Different was the situation of the target feature, Survey counted client, which showed a skewed distribution, but more scattered, similar to a Poisson distribution. This could cause a problem in running a linear regression and more likely we have to run a generalised linear model with a Poisson distribution. This is not surprising, since we are dealing with count data (Zuur et al. 2009). Feature times had as well a skewed distribution, suggesting that the majority of the lectures were concentrating during the early morning and they were decreasing towards the afternoon. ### Box plots.



From the boxplots, all the trends observed in the histograms were confirmed. For categorical variables we plotted bar plot graphs. ### Bar plots.



From the barplots, we could see that observations were equally distributed across all the levels of the feature Room and Factor Time. On the contrary, there were more observations for non lectures and level 3 courses. No issues were detected for those features.

Summary.

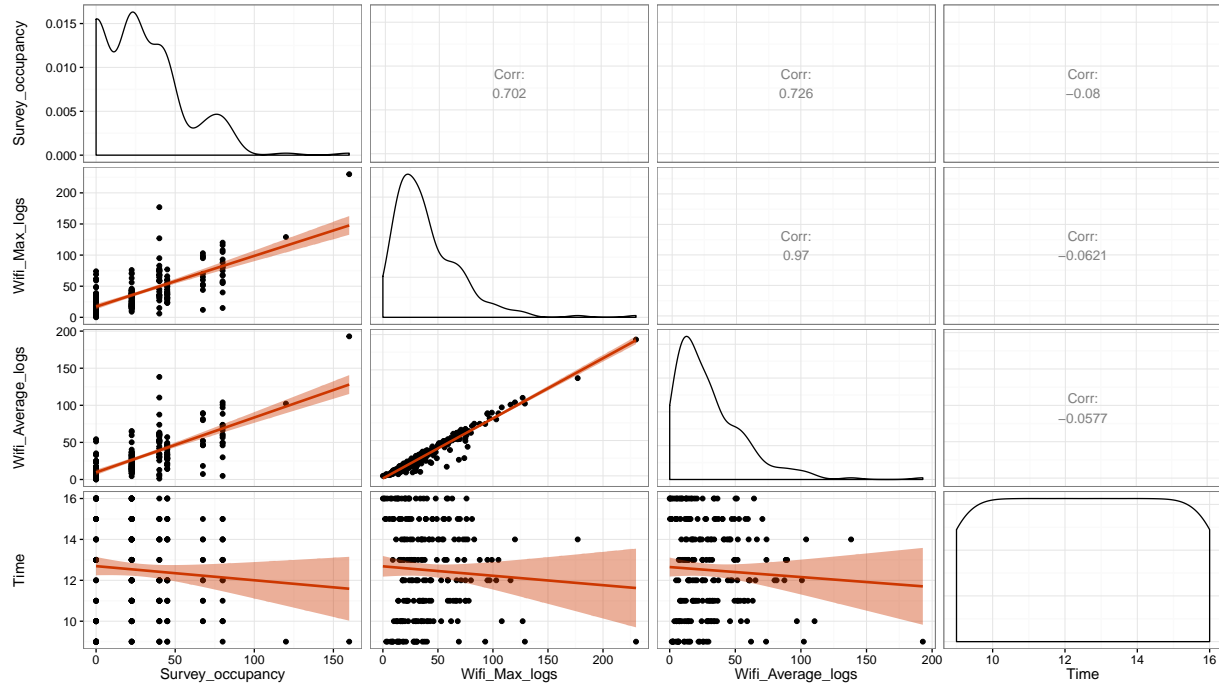
Features	Issues	Planned Solution
Room	None	None
Time	Distribution skewed to the left	To solve during analysis
Factor Time	None	None
Course level	None	None
Tutorial	Uneven representation of the level	Discarded from the analysis
Double Module	Uneven representation of the level	Discarded from the analysis
Class went ahead	Uneven representation of the level	Discarded from the analysis
Wifi Average clients	Distribution skewed to the left & outliers	To solve during analysis
Wifi Maximum clients	Distribution skewed to the left & outliers	To solve during analysis
Survey Counted clients	Distribution skewed to the left & outliers	To solve during analysis

FEATURES AFFECTING THE TARGET FEATURE

The next step of the analysis was to see which features were more likely determining the occupancy of the class.

For the continuous features we explored the effects on the target features using a correlation matrix, while for the categorical features we used box plots.

Correlation matrix for continuous variables.



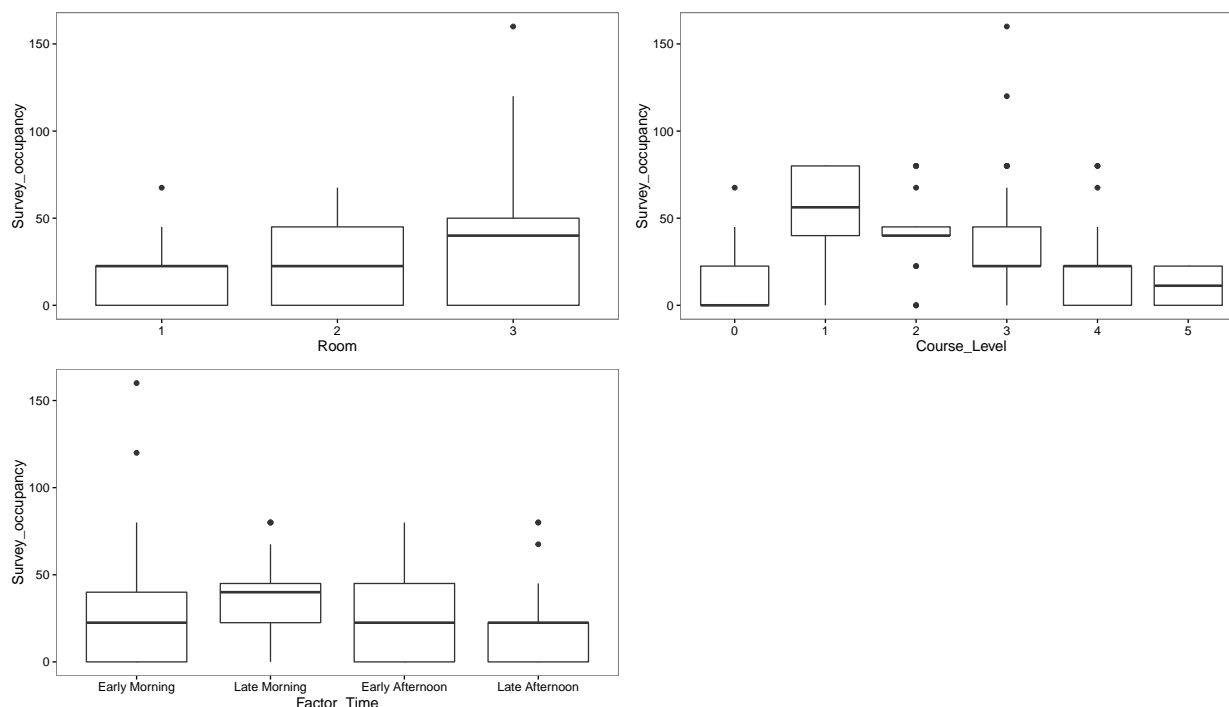
From the correlation matrix Survey countedclients seems to have a good correlation with Wifi Average counted clients and Maximum counted clients, therefore we will try to run 2 models: one for exploring the relationship between Survey counted clients and Average counted clients and another for Survey counted clients and Maximum counted clients. However, from this graph we can see that there are 2 point that are clearly two outliers. Therefore, we are going to run the analyses with and without them to see if there is an improvement of the analysis without them.

From the graphs we can see that Average counted clients and Maximum counted clients are highly correlated showing that both of them are not so different. Therefore we will not expect too much difference among the

2 models.

Time does not seem to be correlated with the target features Survey counted clients and it seems more categorical.

Box plots for categorical variables.



From the boxplot plotting the counted people in the different room, it can be observed that the average of the counted people in room 1 is not different from room2. Room3 has an higher number of people on average, but this can be due to outliers. The average number of counted people, instead, changed across the different course levels and it will be worth to explore if the occupancy of the room was affected by the course level. The highest average number of counted people was around 50 and it was registered in the late morning , while it was around 30 for the rest of the day. Therefore it will be interesting to explore the effect of the time on the occupancy.

Analysis

For selecting the features that together with the max logs or with the average logs best predict the ground truth data we will do model selection with a method similar to the one used by James et al. (2013).

The first step of the analysis is running all the models with all the combinations of the dependent variables, but making sure to always include the Wi-Fi log estimate. The models selected are the following:

- $Survey\ occupancy \sim 1$, the null model;
- $Survey\ occupancy \sim Average\ Wifi\ occupancy$;
- $Survey\ occupancy \sim Average\ Wifi\ occupancy + Room$;
- $Survey\ occupancy \sim Average\ Wifi\ occupancy + Time$;
- $Survey\ occupancy \sim Average\ Wifi\ occupancy + Course_Level$;
- $Survey\ occupancy \sim Average\ Wifi\ occupancy + Room + Time$;
- $Survey\ occupancy \sim Average\ Wifi\ occupancy + Room + Course_Level$;

- *Survey occupancy ~ Average Wifi occupancy + Time + Course_Level;*
- *Survey occupancy ~ Average Wifi occupancy + Room + Time + Course_Level, ;*

The same models will be run also with the Maximum number of logs as dependent variable. All these models were run using the k-fold cross validation. K-fold validation was preferred over the validation set approach and the Leave Out Cross Validation (LOOCV), because it is more robust and more accurate in estimating the test error. The Validation set approach tends to give an over estimate of the test error and the test error is dependent on the observations used to calculate it. Furthermore, the LOOCV tends to provide a test error with a high variance, because the folds used for estimating it are correlated among each other. In this analysis we are going to perform a 10-fold cross validation (10-fold CV), which is pretty standard. The 10-fold CV is carried out with the package: CVglm. For each model we extract the overall mean square error(MSE) and we picked as best model the model with the lowest MSE.

Results

CASE1: Wi-Fi Average logs

All the models ran with the response variable Wifi average logs are summarised in the following table showing their MSE.

Models	MSE
Survey_occupancy ~ 1	711.234
Survey_occupancy ~ Wifi_Average_logs	343.184
Survey_occupancy ~ Wifi_Average_logs + Room	353.37
Survey_occupancy ~ Wifi_Average_logs + Factor_Time	354.343
Survey_occupancy ~ Wifi_Average_logs + Course_Level	366.148
Survey_occupancy ~ Wifi_Average_logs + Room + Factor_Time	364.239
Survey_occupancy ~ Wifi_Average_logs + Room + Course_Level	375.71
Survey_occupancy ~ Wifi_Average_logs + Factor_Time + Course_Level	379.286
Survey_occupancy ~ Wifi_Average_logs + Room + Factor_Time + Course_Level	389.011

As you can see from the table the model with the lowest MSE was the model with only the Wifi average logs as response variable.

CASE 2: Wi-Fi Maximum logs

We run the same models with the max WiFi logs as response variable, in order to see if it was a better predictor than the average WiFi logs.

Models	MSE
Survey_occupancy ~ 1	711.234
Survey_occupancy ~ Wifi_Max_logs	372.782
Survey_occupancy ~ Wifi_Max_logs + Room	384.457
Survey_occupancy ~ Wifi_Max_logs + Factor_Time	385.794
Survey_occupancy ~ Wifi_Max_logs + Course_Level	391.666
Survey_occupancy ~ Wifi_Max_logs + Room + Factor_Time	397.118
Survey_occupancy ~ Wifi_Max_logs + Room + Course_Level	400.692
Survey_occupancy ~ Wifi_Max_logs + Factor_Time + Course_Level	406.344
Survey_occupancy ~ Wifi_Max_logs + Room + Factor_Time + Course_Level	415.618

The model with the lowest MSE is the model: `Survey_occupancy ~ Wifi_Max_logs`. However, its MSE is slightly higher than the previous best model. Therefore the best model was: `Survey_occupancy ~ Wifi_Average_logs`, which we are going to run on the whole dataset.

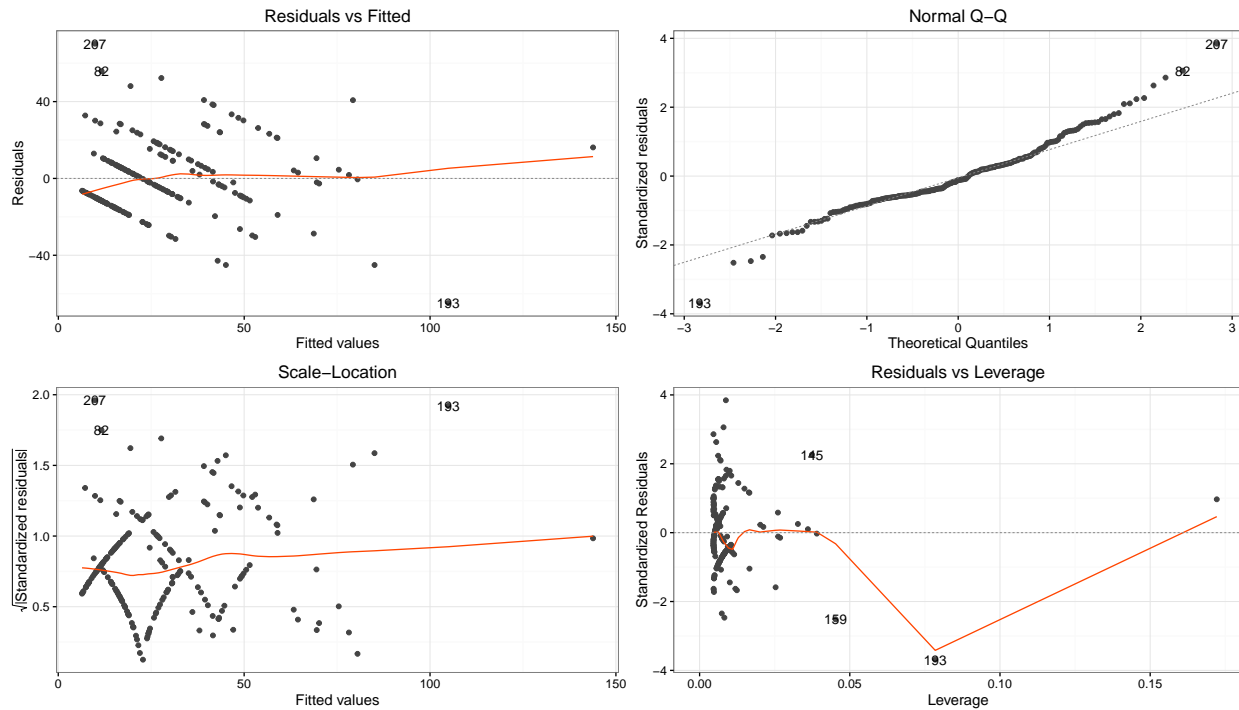
Looking at the model summary, the `Wifi_Average_logs` is significantly related to the ground truth data.

```
summary(occupancy.lm.avg)
```

```
##
## Call:
## lm(formula = Survey_occupancy ~ Wifi_Average_logs, data = AnalysisTable)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -64.90 -10.97  -2.13   9.17  70.16
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6.3992     1.8744   3.41 0.00077 ***
## Wifi_Average_logs  0.7125     0.0461  15.44 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.3 on 214 degrees of freedom
## Multiple R-squared:  0.527, Adjusted R-squared:  0.525
## F-statistic: 238 on 1 and 214 DF, p-value: <2e-16
```

However when we look at the residuals plotted, there are a few issues. As it can be seen below from the plot, showing the fitted values plotted against the residuals, the target features had a lot of values close together, similarly to what was expected from the categorical features and there were potential outliers (fitted values > 140). The observations seemed normally distributed, but the variance did not seem homogeneous.

```
## [1] "ggmultiplot"
## attr(,"package")
## [1] "ggfortify"
```



Removing the outliers

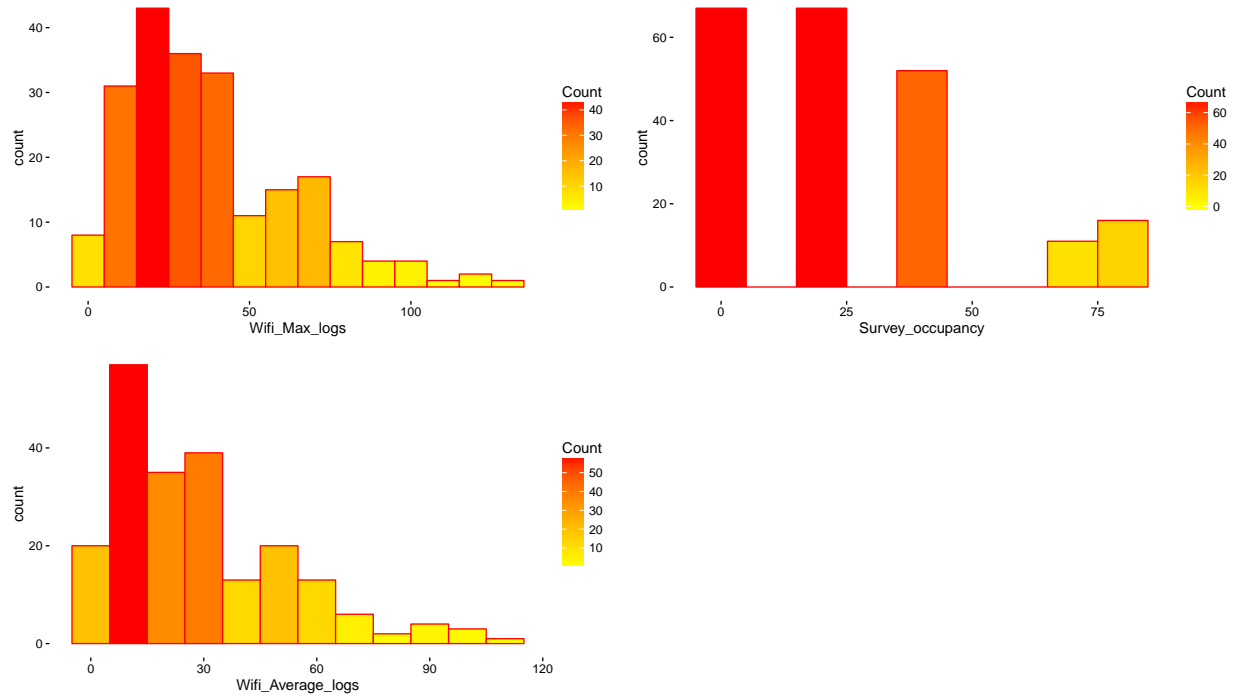
For deciding which observation to remove, we looked at the histograms (see above), which showed that for Wi-fi Average logs and Maximum logs had potential outliers when the values were higher than 140, while for the Survey occupancy observations higher than 120 seemed outliers. Only 3 observations were removed and we did not lose too much data.

```
## [1] 213 14
```

```
## Room      Date      Time      Module      Course_Level
## 1:72 Length:213 Min. : 9.0 Length:213 0:59
## 2:72 Class :character 1st Qu.:11.0 Class :character 1:13
## 3:69 Mode :character Median :13.0 Mode :character 2:23
## Mean :12.5 3:74
## 3rd Qu.:15.0 4:40
## Max. :16.0 5: 4
## Tutorial Double_module Class_went_ahead Capacity
## 0:207 0:209 0: 22 Min. : 90
## 1: 6 1: 4 1:191 1st Qu.: 90
## Median : 90
## Mean :113
## 3rd Qu.:160
## Max. :160
## Percentage_room_full Wifi_Average_logs Wifi_Max_logs Survey_occupancy
## Min. :0.000 Min. : 0.0 Min. : 0.0 Min. : 0.0
## 1st Qu.:0.000 1st Qu.: 11.5 1st Qu.: 18.0 1st Qu.: 0.0
## Median :0.250 Median : 22.9 Median : 32.0 Median :22.5
## Mean :0.244 Mean : 28.7 Mean : 38.1 Mean :26.9
## 3rd Qu.:0.250 3rd Qu.: 41.2 3rd Qu.: 54.0 3rd Qu.:40.0
```

```
## Max.      :0.750      Max.      :110.4      Max.      :127.0      Max.      :80.0
##          Factor_Time
## Early Morning :52
## Late Morning  :54
## Early Afternoon:53
## Late Afternoon :54
##
##
```

The histograms were re-plotted to see whether there was an improvement.



The histograms for the Wi-Fi logs seemed improved, while Survey data were still scattered similar to a Poisson distribution. We decided to run the linear regression and see if there was any improvements.

CASE 1: Model selection with the dependent variable Wi-Fi average logs without outliers

All the models ran with the dependent variable Wifi average logs are summarised in the following table showing their MSE.

Models	MSE
Survey_occupancy ~ 1	593.138
Survey_occupancy ~ Wifi_Average_logs	314.55
Survey_occupancy ~ Wifi_Average_logs + Room	317.932
Survey_occupancy ~ Wifi_Average_logs + Factor_Time	320.687
Survey_occupancy ~ Wifi_Average_logs + Course_Level	330.378
Survey_occupancy ~ Wifi_Average_logs + Room + Factor_Time	323.554
Survey_occupancy ~ Wifi_Average_logs + Room + Course_Level	333.185
Survey_occupancy ~ Wifi_Average_logs + Factor_Time + Course_Level	336.909
Survey_occupancy ~ Wifi_Average_logs + Room + Factor_Time + Course_Level	339.352

As we can see from the table, the best model was still Survey_occupancy ~ Wifi_Average_logs, with a MSE ~ 315. Removing the outlier improved slightly the fit of the model.

CASE 2: Model selection with the dependent variable Wi-Fi maximum logs without outliers

Models	MSE
Survey_occupancy ~ 1	593.138
Survey_occupancy ~ Wifi_Max_logs	336.665
Survey_occupancy ~ Wifi_Max_logs + Room	340.22
Survey_occupancy ~ Wifi_Max_logs + Factor_Time	344.145
Survey_occupancy ~ Wifi_Max_logs + Course_Level	346.838
Survey_occupancy ~ Wifi_Max_logs + Room + Factor_Time	347.135
Survey_occupancy ~ Wifi_Max_logs + Room + Course_Level	349.223
Survey_occupancy ~ Wifi_Max_logs + Factor_Time + Course_Level	354.54
Survey_occupancy ~ Wifi_Max_logs + Room + Factor_Time + Course_Level	356.273

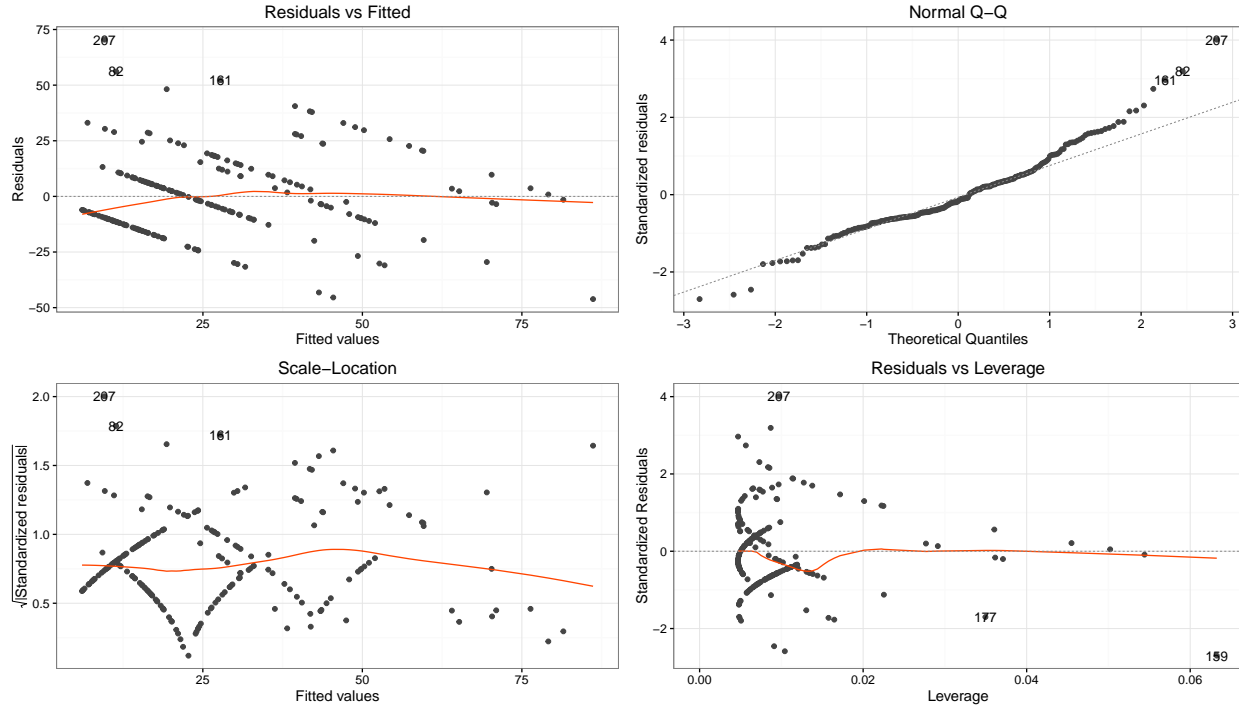
The model with the lowest MSE is the model with only the Wi-Fi Max logs as response variable, whose MSE is slightly higher than the Wi-Fi Average best model. Therefore we are going to consider again the Survey_occupancy ~ Wifi_Average_logs as best model and we run it on the entire data set.

Looking at the model summary, the Wifi_Average_logs were significantly related to the Survey occupancy.

```
##
## Call:
## lm(formula = Survey_occupancy ~ Wifi_Average_logs, data = NoOutlierTable)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -46.16 -10.84  -2.83   8.56  70.43
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.0684     1.9280   3.15  0.0019 **
## Wifi_Average_logs  0.7253     0.0523  13.87 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.6 on 211 degrees of freedom
## Multiple R-squared:  0.477, Adjusted R-squared:  0.475
## F-statistic: 192 on 1 and 211 DF, p-value: <2e-16
```

When we looked at the residuals they were acceptable. From the plot looking at the residual vs fitted values, we still can see that the Survey occupancy was assuming more or less the same values. The data were quite normal distributed, and the variance seems more or less homogeneous.

```
## [1] "ggmultiplot"
## attr(,"package")
## [1] "ggfortify"
```



In order to improve the fit of our linear model we tried to perform a generalised linear model with a Poisson distribution. From previous analysis we saw that this model suffered of overdispersion, therefore corrected the standards errors using a quasi-GLM model, as suggested by Zuur et al. (2009).

GENERALISED LINEAR MODEL WITH POISSON DISTRIBUTION

We ran all the models using the package glm with family quasi-poisson and cv.glm for running 10-fold cross validation and we took the model with the lowest raw cross-validation estimate of prediction (delta) as best model.

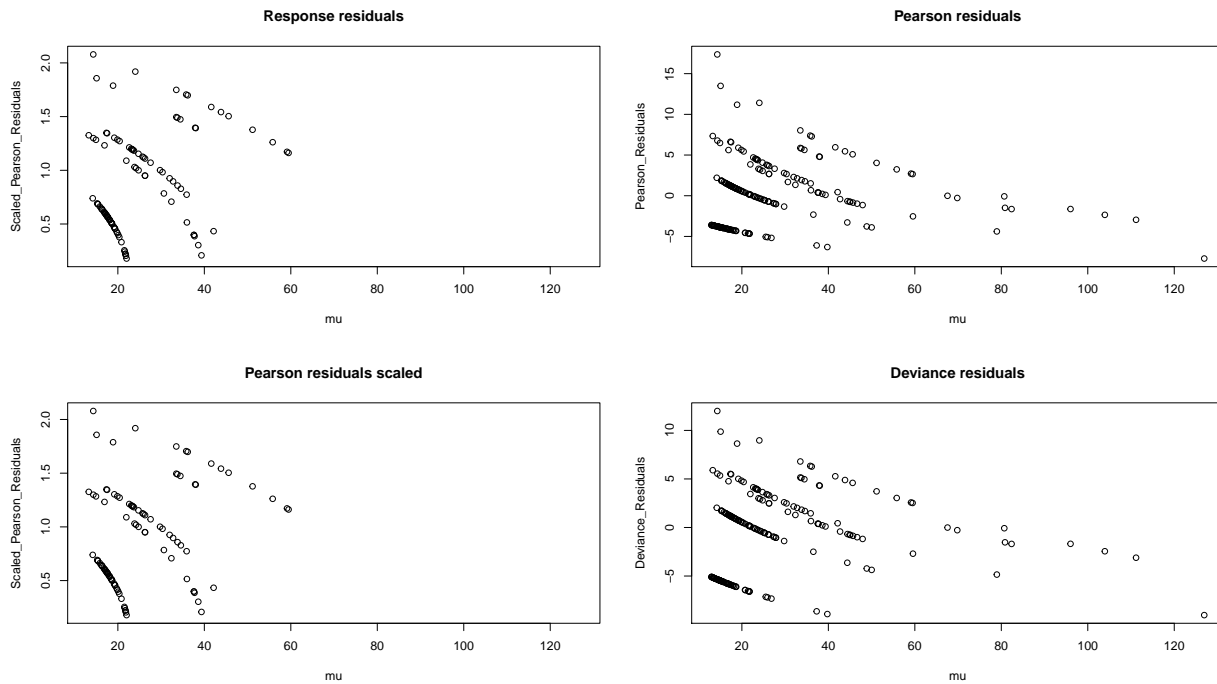
Models	Adjusted delta
Survey_occupancy ~ 1	596.145
Survey_occupancy ~ Wifi_Average_logs	389.82
Survey_occupancy ~ Wifi_Average_logs + Room	395.206
Survey_occupancy ~ Wifi_Average_logs + Factor_Time	390.722
Survey_occupancy ~ Wifi_Average_logs + Course_Level	407.9
Survey_occupancy ~ Wifi_Average_logs + Room + Factor_Time	405.898
Survey_occupancy ~ Wifi_Average_logs + Room + Course_Level	394.589
Survey_occupancy ~ Wifi_Average_logs + Factor_Time + Course_Level	415.519
Survey_occupancy ~ Wifi_Average_logs + Room + Factor_Time + Course_Level	398.368

The best model was Survey_occupancy ~ Wifi_Average_logs with an adjusted cross-validation estimate of prediction error of 390. As for the linear regression we run all the model with the dependent variable Wifi_Max_logs.

Models	Adjusted delta
Survey_occupancy ~ 1	596.145
Survey_occupancy ~ Wifi_Max_logs	393.731

Models	Adjusted delta
Survey_occupancy ~ Wifi_Max_logs + Room	401.427
Survey_occupancy ~ Wifi_Max_logs + Factor_Time	393.774
Survey_occupancy ~ Wifi_Max_logs + Course_Level	404.093
Survey_occupancy ~ Wifi_Max_logs + Room + Factor_Time	411.424
Survey_occupancy ~ Wifi_Max_logs + Room + Course_Level	394.519
Survey_occupancy ~ Wifi_Max_logs + Factor_Time + Course_Level	408.063
Survey_occupancy ~ Wifi_Max_logs + Room + Factor_Time + Course_Level	404.027

The best model was `Survey_occupancy ~ Wifi_Max_logs` with an adjusted cross-validation estimate of prediction error (delta) of: 394, which was slightly worst than the delta of model with the average logs as dependent variable. Therefore we are going to run `Survey_occupancy ~ Wifi_Average_logs` on the whole dataset. To validate the model we plotted the following residuals, as suggested by Zuur et al. (2009): Ordinal residuals, Pearson residual, scaled Pearson residuals (to take into account the overdispersion) and the deviance residuals for the optimal quasi-Poisson model applied on the dataset without outliers. If we spots some patterns in this graphs, it will mean that the distribution selected was not correct or that there is an issue with independence.



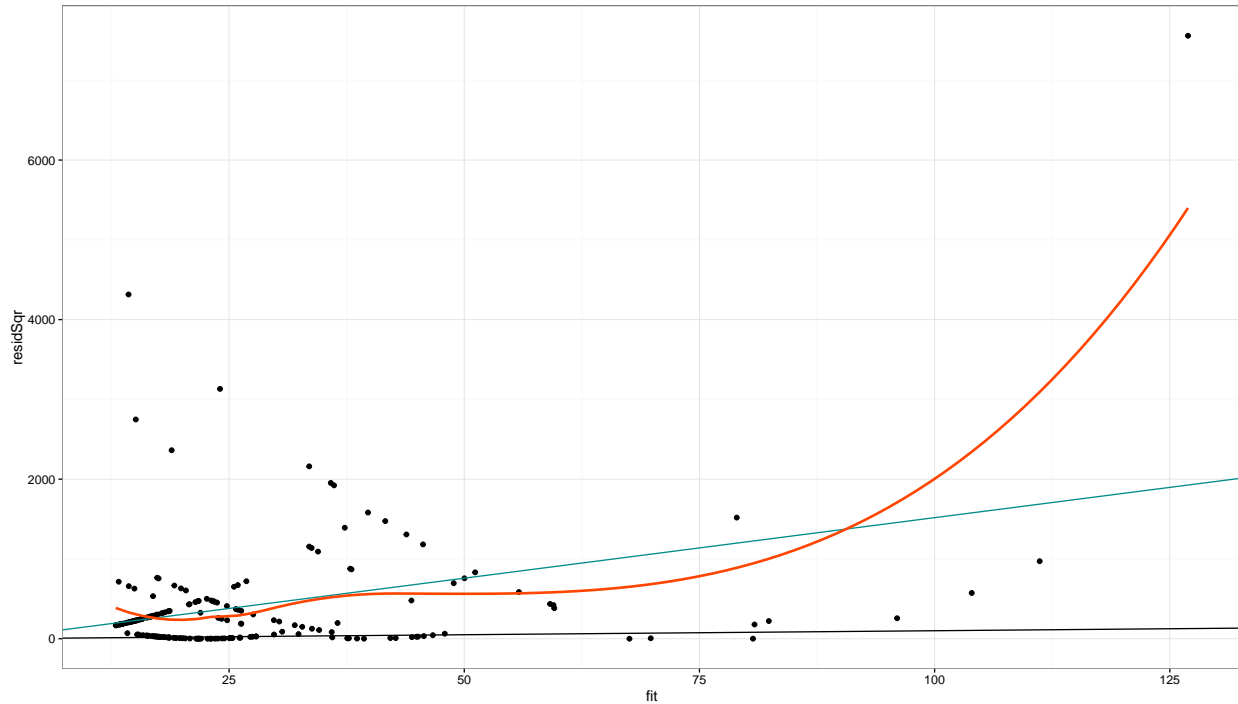
From all the residuals plot we could see a pattern, the residuals were decreasing as the average (μ) of the fitted values were decreasing, suggesting that the quasi-poisson glm was not appropriate.

For double checking it, we checked if the variance of the residuals was proportional to the mean, as suggested by this tutorial ([linked phrase](#)). For doing so we plotted the residuals against the predicted mean and in this graph we plotted 3 lines:

- a black line representing the Poisson assumed variance (~ 1);
- a blue plotting the quasi-Poisson assumed variance;
- orange curve for the smoothed mean of the square of the residual.

In theory the orange line should be straight and collinear with the blue line. Higher is the deviation of

the orange line from the blue one, higher is the chance that the variance of the quasi-Poisson model is not proportional to the mean as assumed by the model.



As we can see from the graph, the orange smoothed mean of the square of the residual is diverging from the Poisson assumed variance. For this reason, the quasi-poisson model was not appropriate for our data. Therefore, we decided to run the model using a negative binomial distribution that it is designed to deal with over dispersion.

The negative binomial model, however, has a very high dispersion parameter and it was not suited for our data.

Therefore, we can conclude that the linear model is our best choice. However, we think that the occupancy data were collected very approximatively. People conducting the survey were only asked to give to the class a score of 0% for empty room, 25% full, 50% full, 75% full or 100% full. These kind of data are more suited for a classification problem. To solve this problem we created a mobile application for helping the people conducting the survey in inputting the right number of people and have a better estimate of the head count in a room. For this reason we still keep the model regression in our application for estimating the count of head in the room.

Reference

Zuur A., Ieno E. & Elphick C.(2009) Mixed Effects Models and Extensions in Ecology with R. Springer.